ED 235 197 TM 830 594

AUTHOR Eignor, Daniel R.; Cook, Linda L.
TITLE An Investigation of the Feasibility of Using Item Response Theory in the Pre-Equating of Aptitude Tests.
INSTITUTION College Entrance Examination Board, New York, N.Y.
PUB DATE Apr 83
NOTE 54p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983). Small print in some figures.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Aptitude Tests; *College Entrance Examinations; *Equated Scores; *Feasibility Studies; Goodness of Fit; *Latent Trait Theory
IDENTIFIERS *Pre Equating (Tests); Scholastic Aptitude Test

ABSTRACT

The purpose of this study was to determine the extent to which item parameters estimated on pretest data from the verbal section of the Scholastic Aptitude Test (SAT) can be used for equating purposes in a situation where intact final form SAT testing data have normally been used. Items appearing in two final SAT-verbal forms were calibrated almost completely from pretest data. An elaborate linkage system was devised and utilized to get parameter estimates for the items, contained in multiple pretests, on the same scale. The final forms under study were equated to two different old forms; the equatings were then redone using item parameter estimates based on the pretest data. For each form, the item response theory (IRT) equating based on pretest statistics was then compared to the IRT equating based on intact final form data and the linear equating used operationally. The results varied considerably across forms, ranging from acceptable to marginally acceptable. The overall results of the pre-equating were deemed sufficiently promising that an investigation of pre-equating two forms of SAT-mathematical will be undertaken. (BW)

# An Investigation of the Feasibility of Using Item Response Theory in the Pre-equating of Aptitude Tests[1,2,3]

Daniel R. Eignor
Linda L. Cook
Educational Testing Service

[1] A paper presented at the annual meeting of AERA, Montreal, 1983.

[2] This study was supported by the College Board through Joint Staff Research and Development Committee funding.

[3] The authors would like to acknowledge the technical assistance of Martha Stocking and Nancy Wright and the programming assistance of Inge Stiebritz and Karen Carroll in performing this study.

# An Investigation of the Feasibility of Using Item Response Theory in the Pre-equating of Aptitude Tests

Daniel R. Eignor
Linda L. Cook
Educational Testing Service

## Introduction

The current thrust of research devoted to the applications of item-response theory (IRT) has generated an active interest in the use of IRT methods in the solution of score equating problems (see Cook and Eignor, 1983). Because of the special properties of test data characterized by IRT models, users are often able to solve problems not amenable to traditional equa~~ting~~ting methods. For other situations, IRT equating offers an alternative against which to evaluate traditional methods. In addition, a number of other important outcomes accrue from the use of IRT for equating tests; among these are: 1) Improved equating, including better equating at the ends of the scale where important decisions are often made, 2) greater test security through less dependence on items in common with a single old form, 3) easier re-equating should items be deleted, and 4) the possible reduction of bias or drift in equating introduced when traditional methods are used over time in certain situations, most notably when the equating samples for the old and new forms are not random samples from the same population.

While the above listed outcomes accrue as the result of the application of any IRT equating method, if the test forms to be equated can be pre-equated using IRT methods, a number of additional advantages accrue. Pre-equating refers to the process of establishing conversions from raw to scaled scores prior to the time the new test is administered operationally. The process depends on the adequate pretesting of a pool of items from which the new test will be built, the calibration of these items using IRT methods, and the utilization of a linking scheme to place the IRT parameters from the pretested items on the same scale. Among the additional advantages offered by IRT pre-equating are: 1) Since equating using IRT pre-equating methods is possible prior to the actual administration of the test, new forms can be introduced at low volume special administrations, a particular problem if traditional methods are used; 2) since pre-equating permits linkages to many old forms, it is the most likely of any equating method to yield acceptable results should testing legislation mandate the disclosure of pretest or equating items; 3) pre-equating would allow more time to do reasonableness and quality control checks, which are normally done in a hurried fashion due to score reporting deadlines; and 4) pre-equating would actually permit a reduction in the usual score reporting cycle while simultaneously allowing more time to do the equating itself. In short, the listed advantages that can potentially accrue from the use of IRT pre-equating build a strong case for investigation of the feasibility of application of this method. In this report, the applicability of IRT pre-equating to the Scholastic Aptitude Test (SAT) verbal section is considered.

## Problem and Purpose

To date, investigations of the feasibility of pre-equating using IRT for tests developed and administered by Educational Testing Service for the College Board have been done using data from the Test of Standard Written English (TSWE) (Bejar and Wingersky, 1982). The Bejar and Wingersky study (1982) indicated some discrepancies between pre-equating results and the results from traditional equatings in situations where traditional equating was a reasonable procedure. The calibration system used for pre-equating TSWE was considerably different, however, from any system that could be devised for pre-equating the SAT. Thus, although the results of the TSWE pre-equating study were not altogether promising, there is little reason to suggest that these results are generalizable to pre-equating the SAT. For this reason, it was deemed important to investigate the feasibility of pre-equating the SAT using an appropriate calibration system, such as that devised for this study.

The purpose of this study was to determine the extent to which item parameters estimated on SAT-verbal pretest data can be used for equating purposes in a situation where intact final form SAT testing data has normally been used. The items that appear in any final SAT-verbal form come from multiple pretests and to the extent that the item parameter estimates are sensitive to the context in which the item appears, or to sample differences, there may be differences between these parameter estimates and parameter estimates generated using data from the actual final form administration, resulting in a discrepancy between equating based on pretest item parameter estimates and intact final form item parameter estimates. More specifically, in the study, items appearing in two final SAT-verbal forms, 3ASA3 and 3BSA3, were calibrated almost completely from pretest data. (See section on IRT Calibration Design and Linkage System.) An elaborate linkage system, quite representative of the system that would exist were pre-equating to be considered for operational use, was devised and utilized to get parameter estimates for the items, contained in multiple pretests, on the same scale. The two verbal forms under consideration were both part of this linkage system.

The effects of using the parameter estimates, obtained from the pretest data, on the equating process were evaluated in the following way. Each of the SAT-verbal final forms under study, when administered for the first time operationally, was equated to two different old forms and the results of the equatings averaged. Conventional linear equating methods were used when this equating was done. These equatings were redone using item parameter estimates based on the pretest data and item parameter estimates generated from the intact final form administration. In each case, IRT true-score equating was performed. For each form, the IRT equating based on pretest statistics was then compared to the IRT equating based on intact final form data and the linear equating used operationally when each form was put on scale. IRT equating based on intact final form data and linear equating results were used as criteria in this study for the following reasons: (1) In recent IRT equating feasibility studies (Petersen, Cook, and Stocking, in press; Kingston and Dorans, 1982), it has been demonstrated that intact form IRT true-score equating is a viable equating method for aptitude test data; and, (2) the linear methods actually performed to put the forms on scale operationally have undergone many years of scrutiny through their use for operational score reporting purposes. This study was done using two SAT-verbal forms so that all results could be replicated. This should form the basis for drawing stronger conclusions about the feasibility of pre-equating the SAT-verbal section than had the replication not taken place.

## Methodology

### Description of Tests

Test booklets containing SAT forms such as those used in this study consist of six 30-minute sections: two SAT-verbal sections, two SAT-mathematical sections, one Test of Standard Written English (TSWE), and one variable section. All examinees at a given administration take the same test sections except for the variable section, where different subsamples of the total group receive different variable sections. The variable section consists of either one of two verbal or mathematical common item equating sections (anchor tests) or one of a number of verbal, mathematical, or TSWE pretests. In this study, data from only the verbal sections, verbal common item equating sections, and verbal pretests were used. The samples used for calibration purposes in the study either took the verbal sections and one of the verbal common item equating sections or the verbal sections and one of the verbal pretests.

The two SAT-verbal sections contain a total of 85 five-choice items (45 items in one section, 40 items in the other section) comprised of 25 antonyms, 20 analogies, 15 sentence completions, and 5 reading passages each of which is followed by 5 items based on the passage. The verbal common item equating sections contain 40 items (10 of each type); these sections are built to be as parallel as possible to the 40 item SAT-verbal section. The verbal pretest sections either contain 45 or 40 items and are built to be as parallel as possible to the comparable length SAT-verbal sections.

Prior to 1982, raw scores on the SAT were typically transformed to scaled scores on the College Board 200 to 800 scale via linear equating methods. Since January of 1982, IRT true-score equating using intact final form data has been used to put forms on scale. SAT-verbal raw scores are obtained scores that have been corrected for guessing. Raw scores are computed by the formula $R-\frac{1}{k}W$, where R is the number of correct responses, W is the number of incorrect responses, and (k+1) equals the number of choices per item.

Item Calibration Design and Linkage System

Pretest items corresponding to the verbal sections of two forms of the SAT, 3ASA3 and 3BSA3, were calibrated and placed on a common scale through an elaborate linkage system which utilized data on overlapping items from the administration of intact final forms with either pretest sections or common item equating sections. The calibration linkage system, involving the pretests, final forms, and equating sections is depicted in Figure 1. Responses from randomly selected samples of approximately 3000 examinees taking each pretest-final form combination and approximately 2700 taking each final form-equating section combination were used for calibration purposes. Each box in Figure 1 represents a separate calibration (computer run). The dotted-lire boxes within the larger boxes indicate the overlapping items that were used to place parameter estimates on the same scale within a single calibration run. The directional arrows between the boxes indicate that a scaling program (described in a later section of this paper) was run to place parameter estimates from the separate calibration runs on the same scale. It should be noted that all items contained in each 40 item equating section appearing in Figure 1 were calibrated; however this was not the case for all items in each pretest or final form. In order to reduce calibration costs, only the 40 item section of SAT-verbal forms used for linking purposes and only the 170[1] (85 items X 2 forms) pretest items which eventually appeared in final forms 3ASA3 and 3BSA3 were calibrated. Table 1 contains the total number of items and also the total number of examinees responding to the items for each of the 13 calibration runs. Table 2 lists the number of pretest items calibrated in each of the runs. Further reduction in costs were made possible by using existing parameter estimates from the SAT IRT Scale Drift Study (Petersen, Cook, and Stocking, in press) whenever possible. Also, certain final form-equating section combinations from the Scale Drift Study (labeled C-G in Figure 1) and certain final form-equating section calibration runs (numbered 9 and 13 in Figure 1) were linked into the overall calibration linking system though they were not essential to getting the pretest parameter estimates on the same scale. This was done for equating purposes, and will be described in a later section.

## IRT Model and Item Calibration

Item response theory (IRT) assumes that there is a mathematical function which relates the probability of a correct response on an item

[1] Pretest data did not exist for 8 of the 85 items in Form 3ASA3. Therefore, final form data had to be used in the calibration system. This data was obtained from calibration run number 9 in Figure 1.

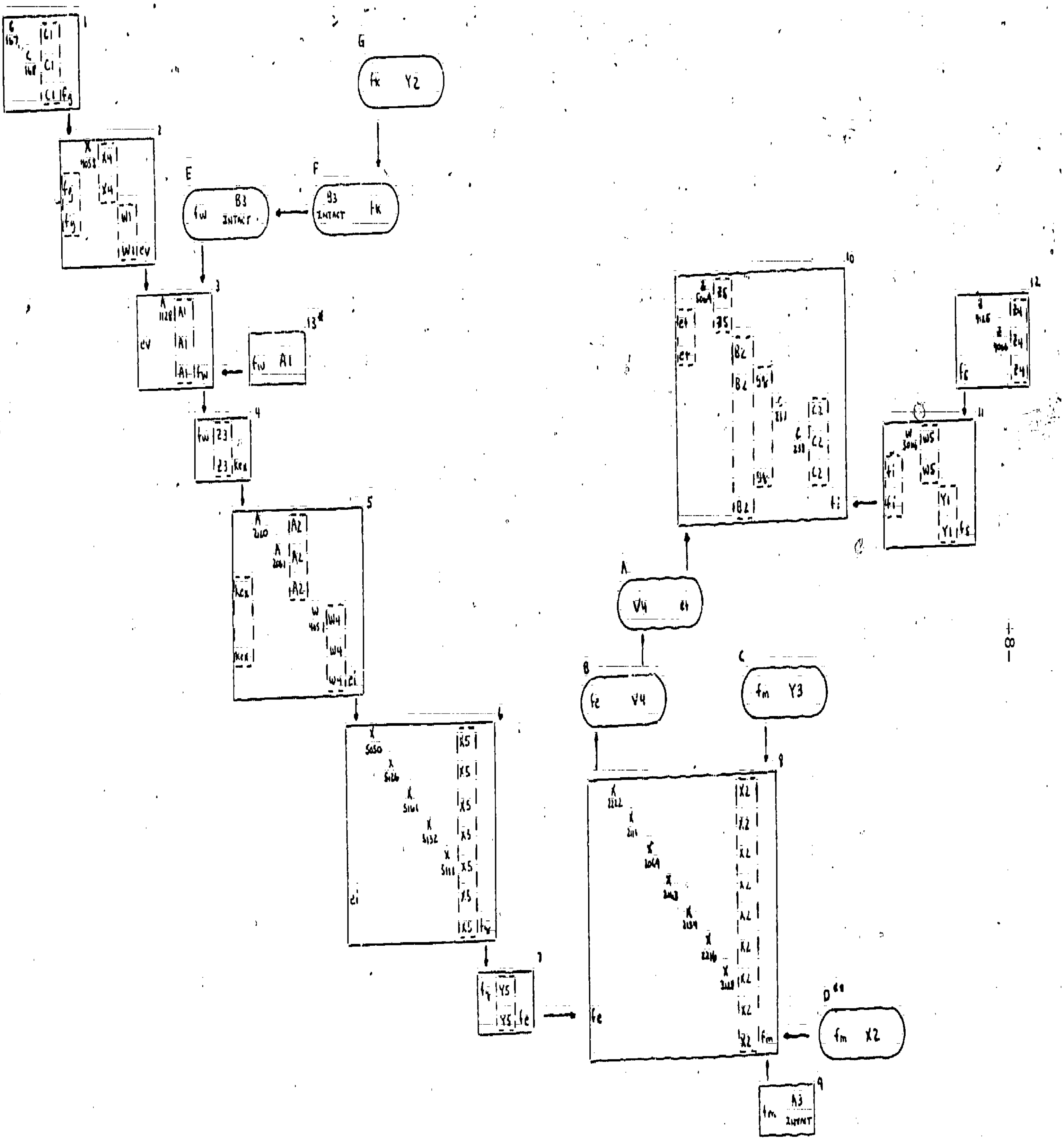


Figure 1: Verbal pre-equating calibration and linking plan. Upper-case letters followed by one digit designate intact SAT final forms. (In the diagram, all intact SAT final form designations are abbreviated; A3 is an abbreviation for 3ASA3, etc.) Upper-case letters followed by three or four digits designate pretest forms. Lower-case letters designate common item equating sections. Boxes (numbered 1-13) indicate separate LOGIST calibration runs. Dotted boxes within the larger boxes indicate overlapping items that were used to place parameter estimates on the same scale within a LOGIST run. Ovals (lettered A-G) indicate forms and equating sections for which parameter estimates already exist from SAT IRT Drift Study. Arrows indicate direction in which scaling (linking) runs took place.

*LOGIST run number 13 necessary so that all 85 A1 items can be calibrated. (Only 40 items from A1 were calibrated in LOGIST run number 3.)

**Linking of D necessary so that the 45 X2 items not calibrated in LOGIST run number 8 could be placed on scale.

Table 1

Total Number of Items and Total Number of Examinees for each of the LOGIST Calibration Runs

| LOGIST Calibration Run Number[1] | Total Number of Items Calibrated | Number of Pretest Items Calibrated | Number of Equating Section Items Calibrated | Number of SAT-verbal Section Items Calibrated | Total Numbers of Examinees |
|---|---|---|---|---|---|
| 1 | 135 | 55 | 40 | 40 | 8,459 |
| 2 | 162 | 2 | 80 | 80 | 8,519 |
| 3 | 121 | 1 | 80 | 40 | 7,964 |
| 4 | 120 | - | 80 | 40 | 6,181 |
| 5 | 174 | 14 | 80 | 80 | 14,069 |
| 6 | 132 | 12 | 80 | 40 | 22,922 |
| 7 | 120 | - | 80 | 40 | 5,123 |
| 8 | 137 | 17 | 80 | 40 | 25,778 |
| 9 | 125 | - | 40 | 85 | 2,777 |
| 10 | 298 | 58 | 120 | 120 | 20,460 |
| 11 | 161 | 1 | 80 | 80 | 10,347 |
| 12 | 82 | 2 | 40 | 40 | 8,146 |
| 13 | 125 | - | 40 | 85 | 2,754 |
|  | 1,892 | 162[2] | 920 | 810 | 143,499 |

[1] LOGIST run number refers to identification scheme in Figure 1.

[2] Pretest data did not exist for 8 of the 85 items in 3ASA3, and hence, final form data had to be used for calibration purposes. Thus only 162 of the total 170 pretest items were calibrated.

Table 2

Number of Items Calibrated from Each Pretest Form

| Pretest Form | LOGIST Run No.[1] | Total No. of Items Calibrated | No. of Items in 3ASA3 | No. of Items in 3BSA3 |
|---|---|---|---|---|
| C167 | 1 | 27 | 13 | 14 |
| C168 | 1 | 28 | 16 | 12 |
| X4058 | 2 | 2 | - | 2 |
| A1128 | 3 | 1 | - | 1 |
| A2120 | 5 | 7 | - | 7 |
| A2061 | 5 | 4 | - | 4 |
| W4057 | 5 | 3 | 3 | - |
| X5050 | 6 | 3 | - | 3 |
| X5126 | 6 | 2 | - | 2 |
| X5161 | 6 | 1 | - | 1 |
| X5132 | 6 | 1 | - | 1 |
| X5111 | 6 | 5 | - | 5 |
| X2222 | 8 | 2 | 1 | 1 |
| X2111 | 8 | 1 | - | 1 |
| X2069 | 8 | 1 | - | 1 |
| X2163 | 8 | 2 | 1 | 1 |
| X2134 | 8 | 4 | 4 | - |
| X2216 | 8 | 1 | 1 | - |
| X2128 | 8 | 6 | 6 | - |
| Z5069 | 10 | 1 | - | 1 |
| C237 | 10 | 29 | 15 | 14 |
| C238 | 10 | 28 | 14 | 14 |
| W5014 | 11 | 1 | 1 | - |
| Z4125 | 12 | 1 | 1 | - |
| Z4066 | 12 | 1 | 1 | - |
| Totals | | 162[2] | 77[2] | 85 |

[1] LOGIST run number refers to the identification scheme in Figure 1.

[2] Pretest data did not exist for 8 of the 85 items in 3ASA3, and hence, final form data had to be used for calibration purposes. Thus, only 77 (of 85) pretest items were calibrated for 3ASA3 and 162 (of 170) for both forms.

to an examinee's ability. (See Lord, 1980, for a detailed discussion.) Many different mathematical models of this functional relationship are possible. The model chosen for this study was the three-parameter logistic model. In this model, the probability of a correct response to item i, $P_i(\theta)$, is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.702\, a_i(\theta - b_i)}} , \qquad (1)$$

where $a_i$, $b_i$, and $c_i$ are three parameters describing the item and $\theta$ represents an examinee's ability. These parameters have specific interpretations: $b_i$ is the point on the $\theta$ metric at the inflection point of $P_i(\theta)$ and is interpreted as the item difficulty; $a_i$ is proportional to the slope of $P_i(\theta)$ at the point of inflection and represents the item discrimination; and $c_i$ is the lower asymptote of $P_i(\theta)$ and represents a pseudo-guessing parameter.

The item parameters and examinee abilities for this study were calibrated using the program LOGIST (Wingersky, Barton, and Lord, 1982; Wingersky, 1983). The estimates are obtained by a (modified) maximum likelihood procedure with special procedures for the treatment of omitted items (see Lord, 1974).

LOGIST requires as input the responses to a set of items from a group of examinees, coded to reflect items answered correctly, incorrectly, omitted and not reached. In addition, the user may specify certain restrictions on the data and parameters in order to speed convergence of the iterative procedure. The major restrictions specified for the study for most of the LOGIST computer runs were:

1. examinees who answered less than one-third of the items were not used,
2. a's were restricted to a range of .01 to 1.75,
3. c's were restricted to a range of .0 to .50 or .75(p+), and
4. θ's were restricted to a range of -7.0 to 5.0.

LOGIST produces as output estimates of the a, b, and c for each item, and θ for each examinee.

Thirteen separate LOGIST runs were necessary to calibrate the pretest items, final form and equating section items used for linking purposes, and the final forms to be used for equating purposes. These LOGIST runs are numbered 1-13 in Figure 1. Each of the separate LOGIST runs generated item parameter estimates on the particular scale defined by the ability distribution of the group of examinees used in the calibration, and hence, a scaling program had to be run to put parameter estimates from the separate LOGIST runs on a common scale. This scaling program also had to be run to put the final form-equating section combinations from the SAT IRT Scale Drift Study (Petersen et al, in press) on the common scale. LOGIST run 10 in Figure 1 was chosen as the base form for scaling purposes because it contains an SAT-verbal form and equating section which are in common with a partial pre-calibration linkage system recently devised (Cook and Petersen, 1982) for possible future operational SAT use.

## Scalings

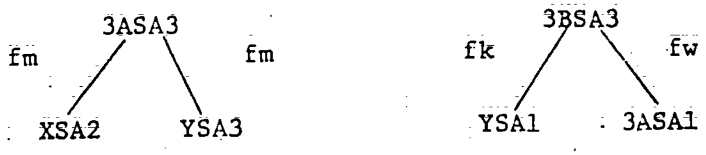
The scalings just referred to are indicated by the directional arrows in Figure 1 (and also Figure 2, to be discussed in the following section). A recently devised scaling method (Stocking and Lord, 1982) was used in the study. Briefly, the method works as follows. Letting b, a, and c denote item difficulty, discrimination, and lower asymptote parameters, a linear transformation of the form

$$\begin{aligned} b_T &= rb + m, \\ a_T &= a/r \qquad (T = \text{transformed}) \end{aligned} \tag{2}$$

is found which places new form item parameters on the base form scale. The r and m of this transformation are chosen to minimize the average squared difference between true scores on the common item set for a particular group of examinees who have taken the base form. It should be noted that $c_T = c$, so that there is no necessity to transform lower asymptote parameters. This method implicitly makes use of data from all the parameters characterizing an item because true scores are used in the minimization process.

Equating Design

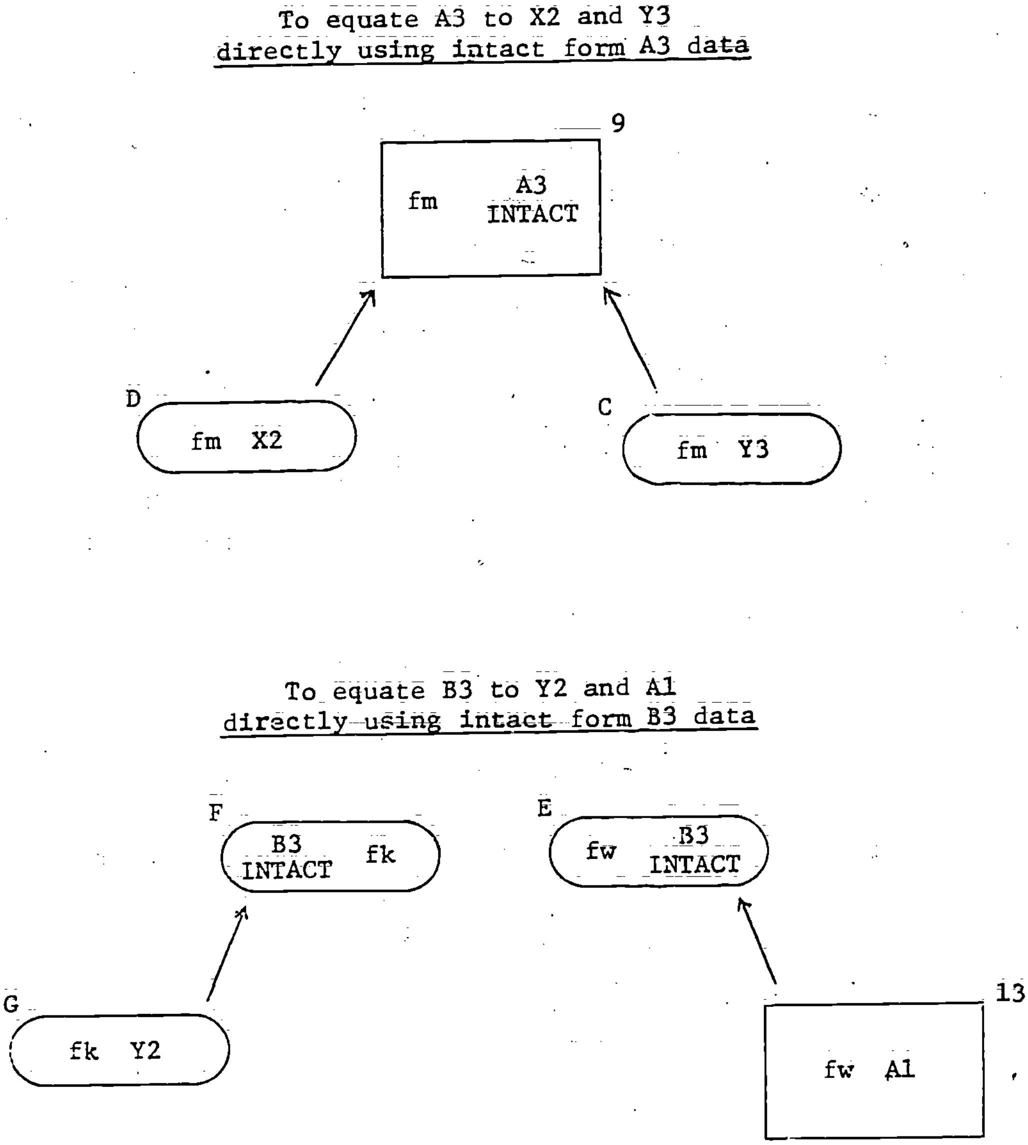
Operationally, the verbal sections of 3ASA3 and 3BSA3 were each linearly equated to two old SAT forms and the results averaged. These equatings can be used as a means for evaluating the effects of using items calibrated from pretest data in the equating process. The following diagram depicts the actual equatings that took place, and the common item sections used for the equatings.



For each equating depicted, IRT true-score equating, to be described in detail in the next section, was done three different ways. The first way, referred to as IRT pre-equating, involved the use of item parameter estimates based on pretest items which constitute 3ASA3 and 3BSA3, while the other two ways (both used as criteria to evaluate the IRT pre-equating) involve the use of item parameter estimates based on data collected when 3ASA3 and 3BSA3 were administered as final forms in an intact fashion. The second and third ways differ in the following fashion. In one situation, referred to as intact form calibration system equating, item parameter estimates for 3ASA3, 3BSA3, and the old forms to which they were equated were placed on the same scale, which is essential for IRT equating, by being linked into the overall calibration and linking plan shown in Figure 1. In this situation, the forms to be equated were linked indirectly through multiple scaling runs applied to a number of intervening LOGIST runs which contain multiple final forms and equating sections. This was done in an attempt to simulate conditions of one possible model under which intact final form IRT equating might take place for the SAT in the future. In the other case, referred to as intact form direct link equating, parameter estimates for the new (3ASA3 and 3BSA3) and old forms to be equated were linked directly through common equating sections. This linking is depicted in Figure 2.

## Equating Methods

Linear equating methods produce an equating transformation of the form $T(x) = Ax + B$, where T is the equating transformation, x is the test score to which it is applied, and A and B are parameters estimated from the data. The Tucker, Levine Equally Reliable, and Levine Unequally Reliable linear equating models (Angoff, 1971, pp. 579-583) are the models that have been used until 1982 for equating SAT-verbal.



To equate A3 to X2 and Y3 directly using intact form A3 data

To equate B3 to Y2 and A1 directly using intact form B3 data

Figure 2: Verbal intact form direct link calibration and linking plan. Upper-case letters followed by one digit designate intact SAT final forms. Lower-case letters designate common item equating sections. Boxes and ovals are numbered to directly correspond to comparable boxes and ovals in Figure 1. Arrows indicate direction in which scaling (linking) took place.

Choice of which of the three models to use for score reporting purposes depends on 1) differences in ability between new and old form groups, as measured by a set of common items (anchor test), and 2) whether the new and old forms are equally reliable, which is typically interpreted to mean of equal test length. These models are based on univariate selection sampling theory. Scores on the relevant selection attribute (the attribute on which the equating samples vary) are assumed to be collinear with scores on the anchor test in the case of the Tucker model and with true scores on both the anchor test and the test form in the case of the Levine models. Scores on the common item set (anchor test) are used to estimate performance of the combined group of examinees on both the old and new forms of the test; thus simulating by statistical methods the situation in which the same group of examinees takes both forms of the test.

The parameters A and B of the equating tranformation are estimated by means of an equation that expresses the relationship between raw scores on two test forms in standard score terms:

$$\frac{X - M_x}{S_x} = \frac{y - M_y}{S_y} , \qquad (3)$$

where x and y refer to the test scores to be equated, and M and S refer to the means and standard deviations of the scores in some group of examinees. Methods using the above equation differ in their identification of the means and standard deviations to be estimated. The Tucker and Levine Equally Reliable methods are based on the estimated means and standard deviations of observed scores whereas the Levine Unequally Reliable method is based on the estimated means and standard deviations of true scores.

## Table 3

### Formulas for Linear Conversion Parameters

**Tucker**

$$A = (S^2_{yb} + C^2_{yvb}(S^2_{vc} - S^2_{vb})/S^4_{vb})^{\frac{1}{2}}(S^2_{xa} + C^2_{xva}(S^2_{vc} - S^2_{va})/S^4_{va})^{-\frac{1}{2}}$$

$$B = M_{yb} + C_{yvb}(M_{vc} - M_{vb})/S^2_{vb} - AM_{xa} - AC_{xva}(M_{vc} - M_{va})/S^2_{va}$$

**Levine Equally Reliable**

$$A = (S^2_{yb} + (S^2_{yb} - S^2_{y''b})(S^2_{vc} - S^2_{vb})/(S^2_{vb} - S^2_{v''b}))^{\frac{1}{2}}$$

$$(S^2_{va} + (S^2_{xa} - S^2_{x''a})(S^2_{vc} - S^2_{va})/(S^2_{va} - S^2_{v''a}))^{-\frac{1}{2}}$$

$$B = M_{yb} + (M_{vc} - M_{vb})((S^2_{yb} - S^2_{y''b})/(S^2_{vb} - S^2_{v''b}))^{\frac{1}{2}}$$

$$- AM_{xa} - A(M_{vc} - M_{va})((S^2_{xa} - S^2_{x''a})/(S^2_{va} - S^2_{v''a}))^{\frac{1}{2}}$$

**Levine Unequally Reliable**

$$A = ((S^2_{yb} - S^2_{y''b})/(S^2_{vb} - S^2_{v''b}))^{\frac{1}{2}}((S^2_{xa} - S^2_{x''a})/(S^2_{va} - S^2_{v''a}))^{-\frac{1}{2}}$$

$$B = M_{yb} + (M_{va} - M_{vb})((S^2_{yb} - S^2_{y''b})/(S^2_{vb} - S^2_{v''b}))^{\frac{1}{2}} - AM_{xa}$$

**Angoff Error Variance Estimates** (Anchor Test External to Total Test)

$$S^2_{p''g} = (S^2_{pg}S^2_{vg} - C^2_{pvg})/(S^2_{vg} + C_{pvg})$$

$$S^2_{v''g} = (S^2_{pg}S^2_{vg} - C^2_{pvg})/(S^2_{pg} + C_{pvg})$$

**Notation**

| | |
|---|---|
| New Test Form | X |
| Old Test Form | Y |
| Either New or Old Test Form | P |
| Anchor Test | V |
| Observed Score | x, y, v, p |
| Error Score | x", y", v", p" |
| Group Taking Test X and Test V | a |
| Group Taking Test Y and Test V | b |
| Group Taking Test P and Test V | g |
| Combined Group | c or (a + b) |
| Mean | M |
| Standard Deviation | S |
| Covariance | C |

The formulas for computing the A and B parameters for the Tucker, Levine Equally Reliable, and Levine Unequally Reliable models are given in Table 3. As noted in Table 3, the formulas for the Levine models require error variance estimates. Angoff's method (1953) of estimating error variances is used for operational linear equating. This method assumes that the test to be equated and the anchor test are parallel except for length.

When a new form is equated to two old forms, the final linear parameters to put the new form on scale are arrived at in the following fashion. Each of the old forms has linear parameters for placing it on scale; these parameters are combined with linear parameters generated from the equating relationship to derive parameters to put the new form on scale. There will be a set of parameters for each equating to each old form; the final set of parameters are arrived at by averaging the parameters from each of the single equatings.

Although there are a number of equating techniques possible when using IRT, this study was concerned only with true formula score equating (Lord, 1980). The expected value of an examinee's observed formula score is defined as his or her true formula score. For the true formula score, $\xi$ , we have

$$\xi = \sum_{i=1}^{n} \left[ \frac{(k_i + 1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right] \qquad (4)$$

where n is the number of items in the test and $(k_i+1)$ is the number of choices for item i. If we have two tests measuring the same ability $\theta$ , then true formula scores $\xi$ and $\eta$ from the two tests are related by the equations

$$\xi = \sum_{i=1}^{n} \left[ \frac{(k_i + 1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right]$$

$$\eta = \sum_{j=1}^{m} \left[ \frac{(k_j + 1)}{k_j} P_j(\theta) - \frac{1}{k_j} \right] \qquad (5)$$

Clearly, for a particular $\theta$ corresponding true scores $\xi$ and $\eta$ have identical meaning. They are said to be equated.

Because true formula scores below the chance score level are undefined for the three-parameter logistic model, some method must be established to obtain a relationship between scores below the chance level on the two test forms to be equated. The approach used for this study (Lord, 1980) was to estimate the mean (M) and standard deviation (S) of below chance level scores on the two tests to be equated via the formulas

$$M = \sum_{i=1}^{n} \left[ c_i (c_i + 1)/k_i - 1/k_i \right] , \quad \text{and} \qquad (6)$$

$$S^2 = \sum_{i=1}^{n} (c_i - c_i^2)\ (k_i + 1)^2/k_i^2 ,$$

where n is the number of items in the test, $(k_i+1)$ is the number of choices for item i, and $c_i$ is the psuedo-guessing parameter for item i; and then to use these estimates to do a simple linear equating (see equation (3)) between the two sets of below chance level scores.

In practice, true score equating is carried out by substituting estimated parameters into the equations (5) and (6). Paired values of $\xi$ and $\eta$ are then computed for a series of arbitrary values of $\theta$. Since we cannot know an examinee's true formula score, we act as if relationships (5) and (6) apply to an examinee's observed formula score.

Two further points require clarification. First, the mechanics of doing IRT true-score equating based on pretest data (pre-equating) and based on intact final form data are exactly the same. What differs are the item parameter estimates that are used to calculate $P_i(\theta)$ in equation (4). In one instance the parameters have been calibrated for the item when given in a pretest, and in the other instance, when the item was given as part of an intact final form. Second, when performing score equating to two old forms using IRT true-score equating techniques, a conversion table is generated for each new form-old form relationship and then the corresponding entries in each table are simply averaged to generate the final table.

## Results

### Pre-equating Results

A number of figures and tables have been prepared to summarize the results of this study. Because the equatings done for 3ASA3 and for 3BSA3 are independent, and meant to serve as replications of the pre-equating process, the figures for each form can be viewed separately. Because each of the forms was equated to two old forms, there are figures for each of the single equatings and then the equating resulting from the averaging of the single equatings.

In the figures for each equating performed, there are two plots. The first plot compares the raw to scaled score conversion line resulting from the IRT pre-equating to one of the three comparison conversion lines, resulting from either the intact form calibration system IRT equating, the intact form direct link IRT equating, or the intact form linear equating actually used operationally for score reporting purposes. The second plot contains residuals. These residuals are simple differences between scaled scores resulting from the IRT pre-equating and one of the comparison equatings for each possible formula score point. The plots use the comparison equating (intact form calibration IRT equating, intact form direct link IRT equating, or intact form linear equating) as the baseline and show differences between the pre-equating equating and the baseline equating results across the formula score scale. As mentioned earlier, the intact form calibration system and direct link IRT equatings were chosen as baseline equatings for these residual plots because this sort of IRT equating has been shown in previous studies to be a viable equating method for SAT data, and provides a good criterion equating against which to evaluate the pre-equating results. The intact form linear equating was also used as a baseline because this was the method actually used to put 3ASA3 and 3BSA3 on scale operationally. Of the three comparison equatings, the intact form direct link equating should provide the best criterion against which to evaluate the pre-equating results in that 1) the relationship between the parameter estimates, for the forms to be equated, from the separate LOGIST runs have not been influenced by intervening scalings, and 2) in contrast with linear

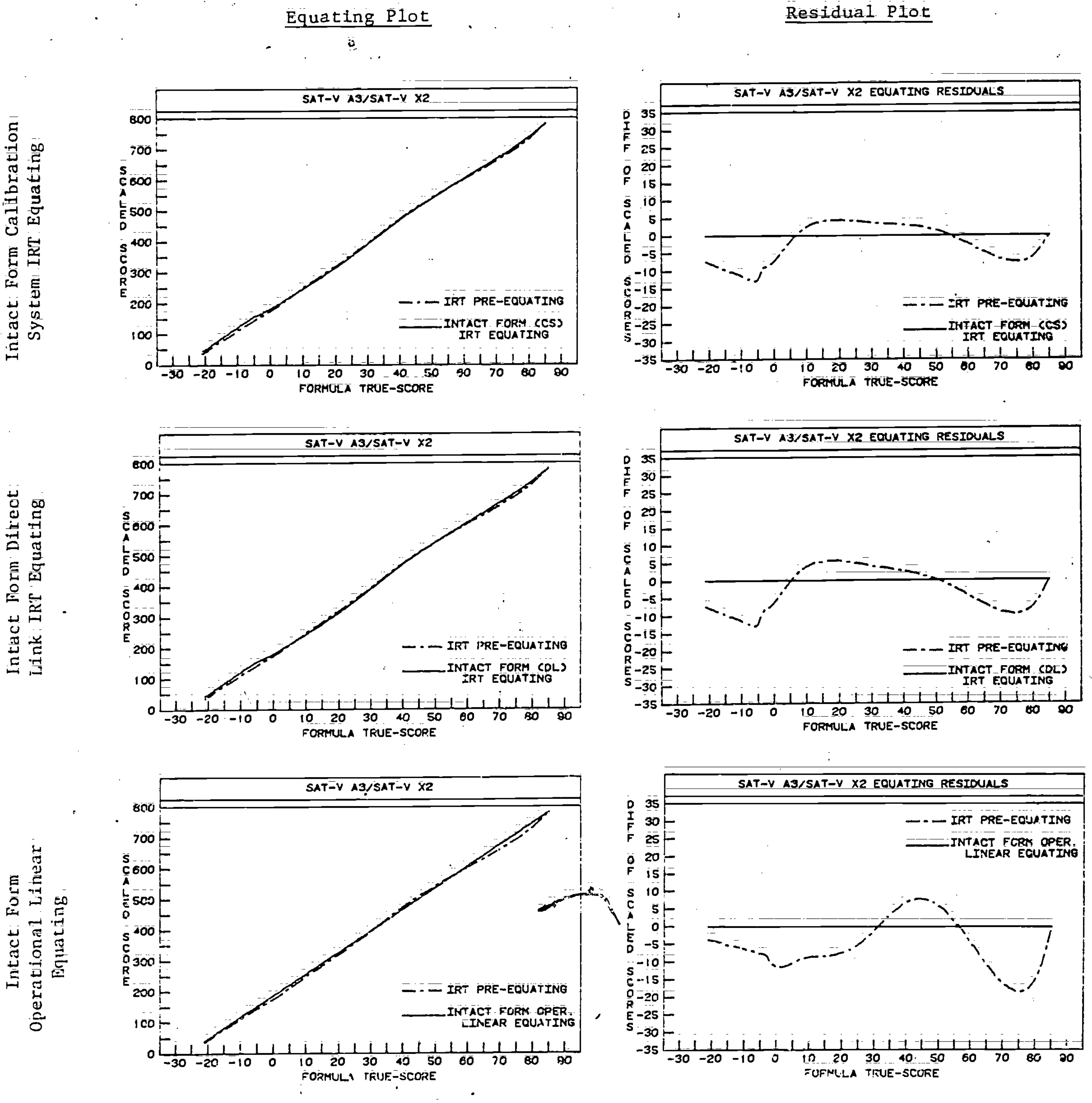


Figure 3: SAT-verbal Form 3ASA3 equated to SAT-verbal Form XSA2 - Plots of 1) IRT pre-equating raw to scaled score transformation compared to corresponding intact form calibration system IRT, direct link IRT, and operational linear equating raw to scaled score transformations, and 2) differences between scaled scores (IRT pre-equating - comparison equating) resulting from the equatings.

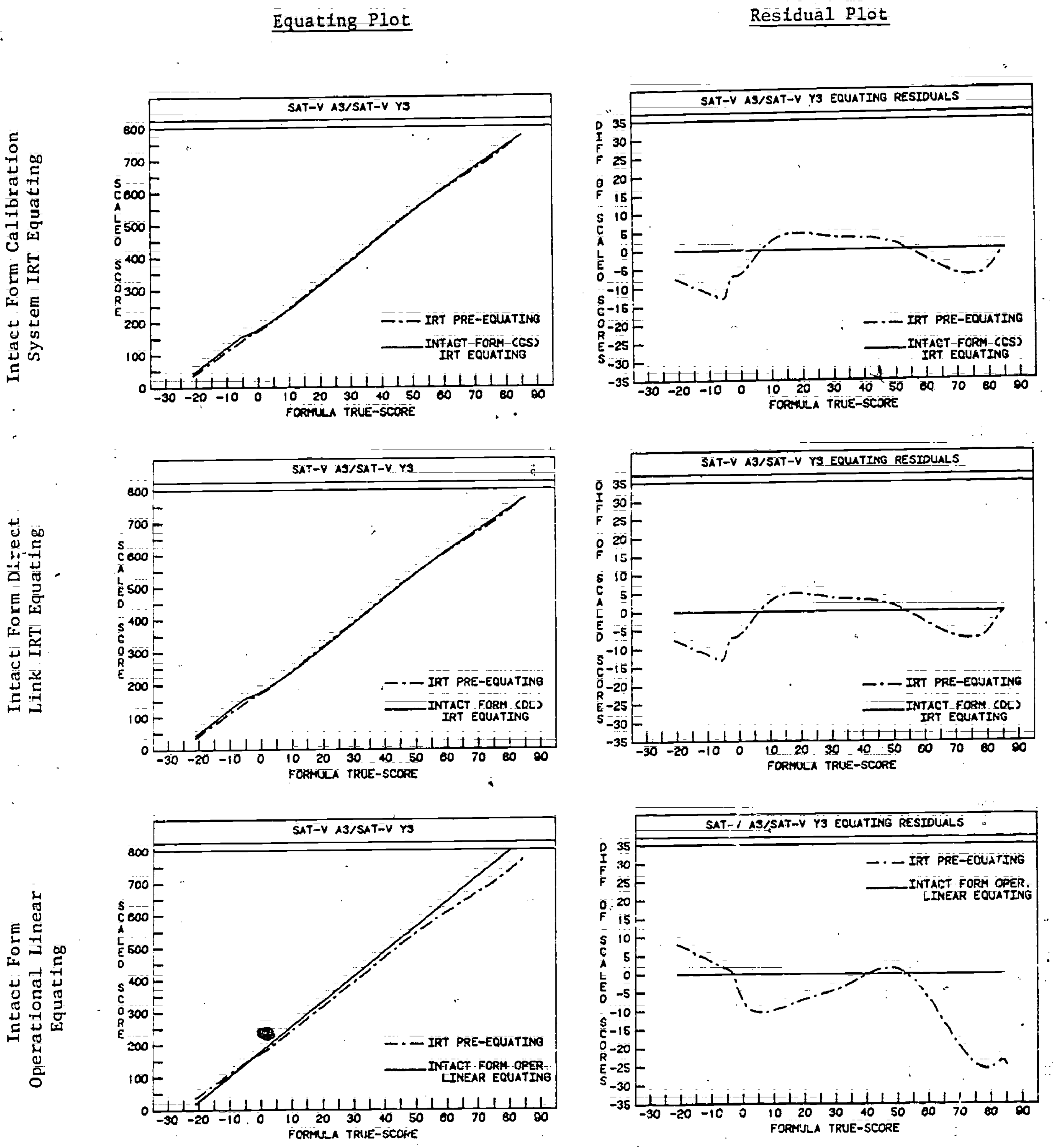


Figure 4: SAT-verbal Form 3ASA3 equated to SAT-verbal Form YSA3 - Plots of 1) IRT pre-equating raw to scaled score transformation compared to corresponding intact form calibration system IRT, direct link IRT, and operational linear equating raw to scaled score transformations, and 2) differences between scaled scores (IRT pre-equating - comparison equating) resulting from the equatings.

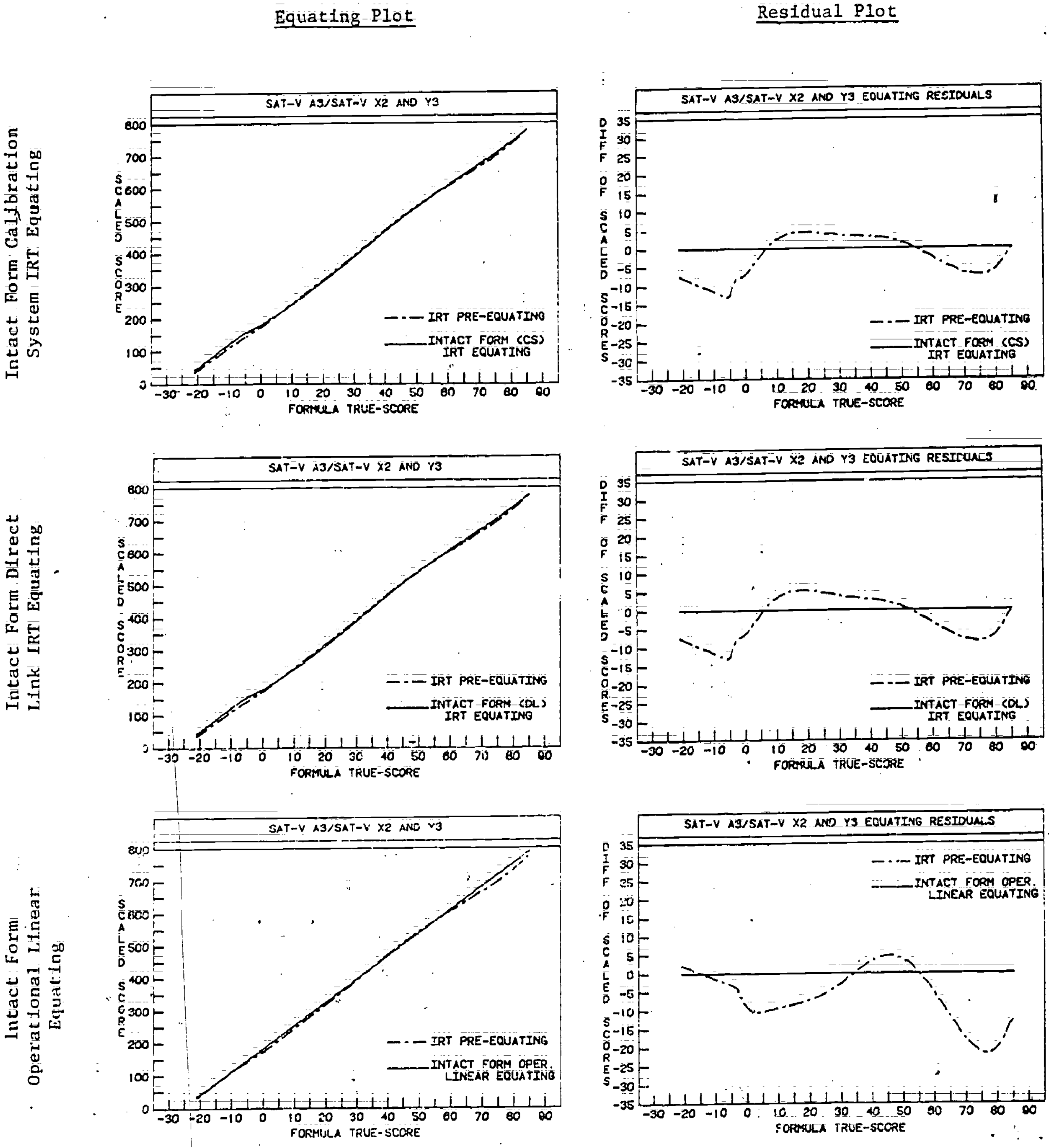


Figure 5: SAT-verbal Form 3ASA3 equated to SAT-verbal Form XSA2 and SAT-verbal Form YSA3 - Plots of 1) IRT pre-equating raw to scaled score transformation compared to corresponding intact form calibration system IRT, direct link IRT, and operational linear equating raw to scaled score transformations, and 2) differences between scaled scores (IRT pre-equating - comparison equating) resulting from the equatings.

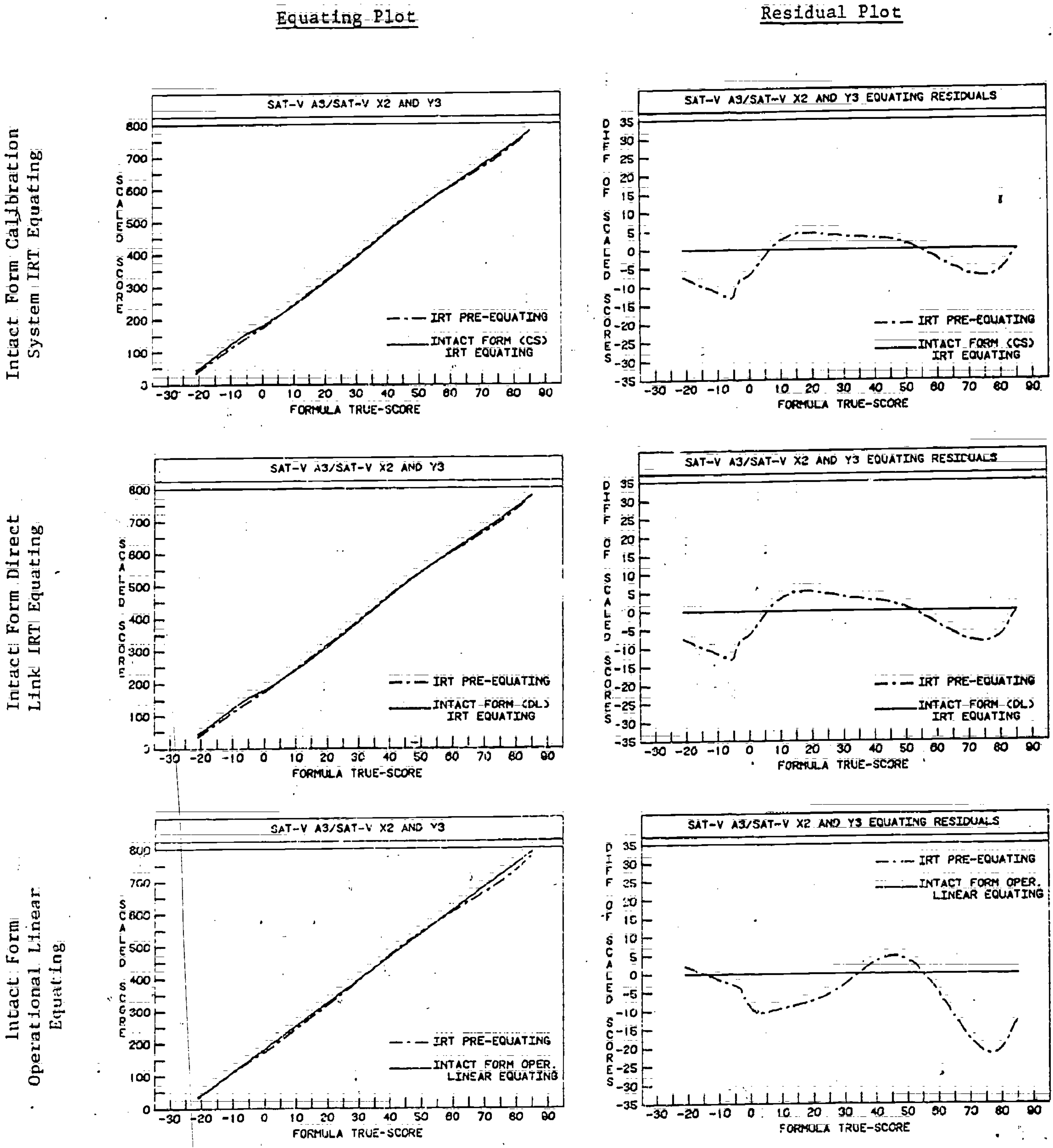


Figure 5: SAT-verbal Form 3ASA3 equated to SAT-verbal Form XSA2 and SAT-verbal Form YSA3 - Plots of 1) IRT pre-equating raw to scaled score transformation compared to corresponding intact form calibration system IRT, direct link IRT, and operational linear equating raw to scaled score transformations, and 2) differences between scaled scores (IRT pre-equating - comparison equating) resulting from the equatings.

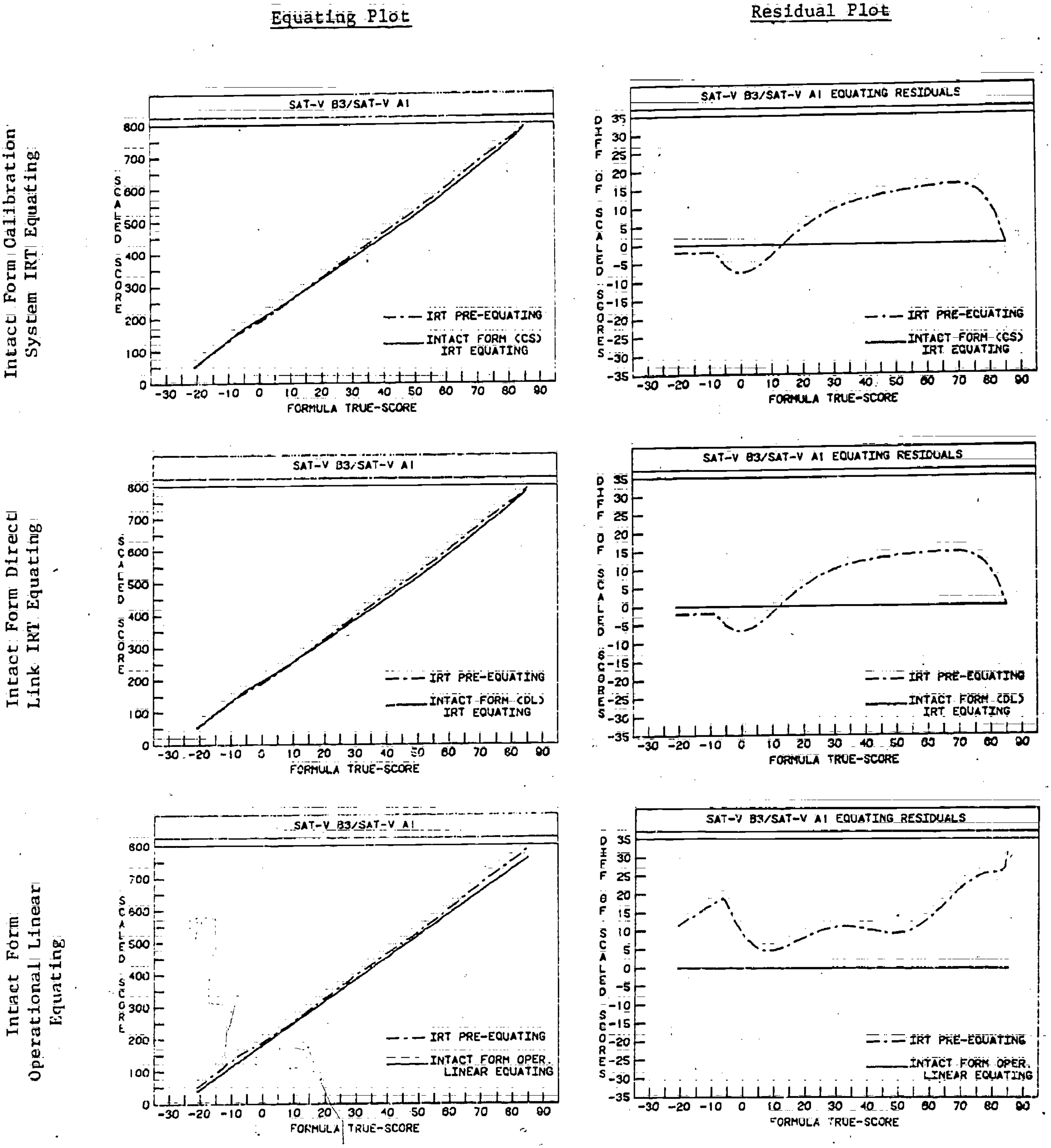


Figure 7: SAT-verbal Form 3BSA3 equated to SAT-verbal Form 3ASA1 - Plots of 1) IRT pre-equating raw to scaled score transformation compared to corresponding intact form calibration system IRT, direct link IRT, and operational linear equating raw to scaled score transformations, and 2) differences between scaled scores (IRT pre-equating - comparison equating) resulting from the equatings.

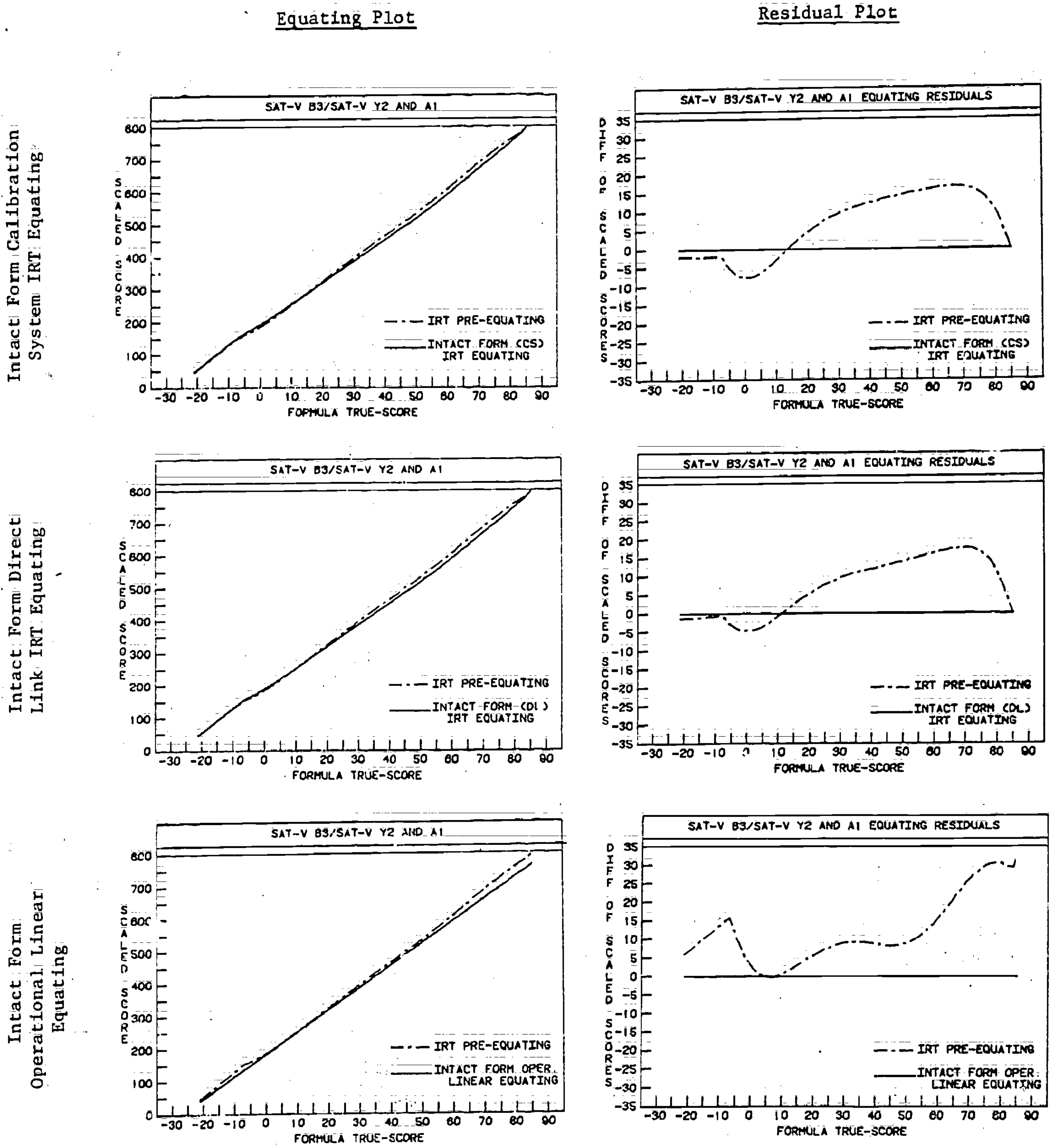


Figure 8: SAT-verbal Form 3BSA3 equated to SAT-verbal Form YSA2 and SAT-verbal Form 3ASA1 - Plots of 1) IRT pre-equating raw to scaled score transformation compared to corresponding intact form calibration system IRT, direct link IRT, and operational linear equating raw to scaled score transformations, 2) differences between scaled scores (IRT pre-equating - comparison equating) resulting from the equatings.

Table 4

Scaled Score Summary Statistics Resulting from Application of Four Equating Methods
SAT-verbal Sections of Forms 3ASA3 and 3BSA3

| Form | N | | IRT Intact Form (Direct Link) | IRT Intact Form (Calibration System) | Intact Form Linear | IRT Pre-equating |
|---|---|---|---|---|---|---|
| 3ASA3 | 126,788 | M | 437.04 | 437.01 | 441.45 | 439.26 |
| | | S.D. | 111.91 | 111.30 | 108.34 | 109.65 |
| 3BSA3 | 253,354 | M | 430.25 | 430.42 | 431.42 | 440.39 |
| | | S.D. | 105.99 | 105.57 | 106.53 | 110.55 |

equating, curvilinear relationships are permitted. The residual plots, in conjunction with data presented in Table 4, to be described next, provide much of the data upon which to evaluate the results of this study.

Table 4 provides the scaled score means and standard deviations for Forms 3ASA3 and 3BSA3 that would have resulted from use for score reporting purposes of pre-equating, intact form calibration system IRT equating, intact form direct link IRT equating, and intact form linear equating to the old forms. The means and standard deviations were computed using frequencies for the total groups taking Forms 3ASA3 and 3BSA3 at the respective initial intact form administrations.

Based on the data presented for Form 3ASA3, it is clear that the pre-equating was quite successful. In no residual plot is the difference between the scaled score resulting from the pre-equating and the comparison intact form calibration system IRT or direct link IRT equatings more than 15 score points on a scale containing 600 score points. The differences between the pre-equating results and the intact form linear results are greater than the differences resulting from the intact form IRT equatings, particularly at the upper end of the formula score scale. This is because all three IRT equatings demonstrate that the raw to scaled score conversion is curvilinear in this region, and the linear equating cannot account for this curvilinearity. The differences in scaled score means and standard deviations presented in Table 4 are very small. The scaled score means and standard deviations resulting from the two IRT methods used as criteria are almost identical. The scaled score mean resulting from the pre-equating lies between the scaled score mean resulting from either the intact form calibration system or direct link IRT equatings and the scaled score mean resulting from the intact form linear equating. The scaled score difference between the mean resulting from the pre-equating and any of the other equatings is about 2 points. What is particularly interesting to note is the pattern of the residuals plots for the comparison of the pre-equating results with the intact form calibration system and direct link IRT equating results, displayed in Figures 3-5. The patterns of residuals are the same across both the single equatings and the average equating. The pre-equating results in lower scaled scores at the bottom and top of the formula score scale and slightly higher scaled scores in the middle region. As mentioned earlier, at no point are these differences greater than 15 scaled score points, and hence, although the pattern of differences is consistent across equatings, the differences themselves are minor when compared to, for instance, the scaled standard error of measurement for SAT-verbal, which is approximately 30 scaled score points.

The pre-equating of Form 3BSA3 was clearly not as successful as the pre-equating for Form 3ASA3. The residual plots show maximum differences in scaled scores resulting from the pre-equating and the operational calibration system or direct link IRT equating of upwards of 20 score points. Once again, the differences between the pre-equating and the intact form linear equating are even greater, particularly in the regions of the formula score scale where the raw to scaled conversion is curvilinear. The differences in scaled score means and standard deviations resulting from the pre-equating and the comparison equatings are much larger than those for Form 3ASA3. The two IRT methods used as criteria produced scaled score summary statistics that are very similar. Unlike the equatings for 3ASA3, scaled score summary statistics produced by the linear equatings are fairly similar to those produced by the IRT criterion equatings. The scaled score mean resulting from the pre-equating is about ten points greater than the scaled score means resulting from the IRT intact form calibration system, IRT intact form direct link, and intact form linear equatings, which are all within a scaled score point of each other. Once again, the patterns in the residual plots for the IRT pre-equating and the comparison IRT equatings are the same across both of the single equatings (to YSA2 and to 3ASA1) and, hence, the subsequent average equating. The pre-equating results in slightly lower scaled scores at the lower end of the formula score scale and consistently higher scaled scores through the middle and upper end of the formula score scale. The maximum differences occur in all plots around a formula score of 70.

Supplemental Investigations and Results

A number of possible explanations were generated for why the 3BSA3 pre-equating results were different from the 3BSA3 comparison equating results and clearly not of the same quality as the 3ASA3 pre-equating results. Exploration of these possible reasons for the inferiority of the 3BSA3 pre-equating results are reported on next, and also discussed briefly in the conclusions section.

One possible explanation for the 3BSA3 pre-equating results has to do with practice effects generated from the manner in which the test sections are sequenced. In other words, for 3ASA3 there may have been more or less of a balancing effect of the sequencing of the operational final form section and the pretest section (perhaps in about 50% of the final form - pretest combinations represented in Figure 1 the operational section appeared first and in the other 50% of the combinations the pretest section appeared first), while for 3BSA3 the balancing may not have occured. It can be hypothesized, given that the 3BSA3 pre-equating results are consistently higher in the upper part of the formula score scale than any of the comparison equatings, that the pretest section followed the operational section in a disproportionate number of cases, and that practice effects resulted. An investigation of the sequencing of sections did indeed show that, for 3BSA3, the pretests occurred after the operational sections in 65% of the final form - pretest combinations, but for 3ASA3, this was true 64% of the time. Hence it would appear that the above explanation cannot be used to explain why the 3BSA3 pre-equating results were so different from those for 3ASA3.

Two other potential explanations for differences in pre-equating results have to do with equating samples and LOGIST calibration runs. These are:

1. The use of two different equating samples with the 3BSA3 intact form calibration system and direct link equatings. In doing the 3ASA3 intact form calibration system and direct link equatings, the same equating section, fm, was in common with old forms XSA2 and YSA3, and hence the same sample, and subsequent set of parameter estimates, could be used for both equatings. This was not true for 3BSA3 in that fk was in common with YSA2 and fw

with 3ASA1. This necessitated the use of two different samples, and hence, two different sets of item parameter estimates (both sets taken from the SAT IRT Scale Drift Study) to perform the equatings.

2. The use of different versions of the LOGIST program to generate item parameter estimates. For 3ASA3, both the pretest and the final form parameter estimates were generated from the current version of LOGIST, and this is also true of the 3BSA3 pretest parameter estimates. To save on calibration costs, the 3BSA3 final intact form parameter estimates were recovered from the SAT IRT Scale Drift Study (Petersen, et al, in press) run on a different version of LOGIST. It is possible that the updating and refinement of the LOGIST program caused subtle differences in parameter estimates, which collectively caused the differences seen in the residual plots for 3BSA3.

The possible explanations above implicitly assume that it is not the pre-equating for 3BSA3 that is somehow faulty, but instead the comparison IRT equatings. To investigate whether or not it is reasonable to explain the differences in pre-equating results this way, the following was done. The operational final form-equating section combinations needed to equate intact final form 3BSA3 to old forms YSA2 and 3ASA1 (see Figure 2) were run together in one large LOGIST run, using the current version of LOGIST and the intact final form equating redone. As a result, the parameter estimates for the 3BSA3 pre-equating and the 3BSA3 intact final form equating were generated using the same version of LOGIST. Further, by running the data for 3BSA3 and the two old forms concurrently, there was no need for scaling parameter estimates (all parameter estimates needed in the equating are automatically on the same scale) and only one set of 3BSA3 final form parameter estimates were used in the equating (unlike the previous IRT comparison equatings). In sum, the results of equating intact final form 3BSA3 to the old forms using the parameter estimates from the concurrent LOGIST run should provide the best criterion possible for evaluating the 3BSA3 pre-equating results.

A comparison of the 3BSA3 pre-equating results to this new IRT comparison equating is presented in Figure 9 for each of the single equatings and the average equating. The new comparison equating has been labeled intact form concurrent equating in this figure. Information on the scaled score summary statistics resulting from this new equating and the others previously described is presented in Table 5.

The results presented in Figure 9 clearly lead to the conclusion that it is not the comparison IRT equatings for Form 3BSA3 that are faulty, but rather the Form 3BSA3 pre-equating results. The data presented in Figure 9 show differences between the IRT pre-equating and the intact form concurrent IRT equating that are comparable to the differences in the residual plots using the other intact form comparison equatings. Thus the possible explanations for differences in equating results based on the use of different versions of LOGIST and multiple sets of parameter estimates, generated from the IRT Scale Drift Study (Petersen et al, in press) must be discounted.

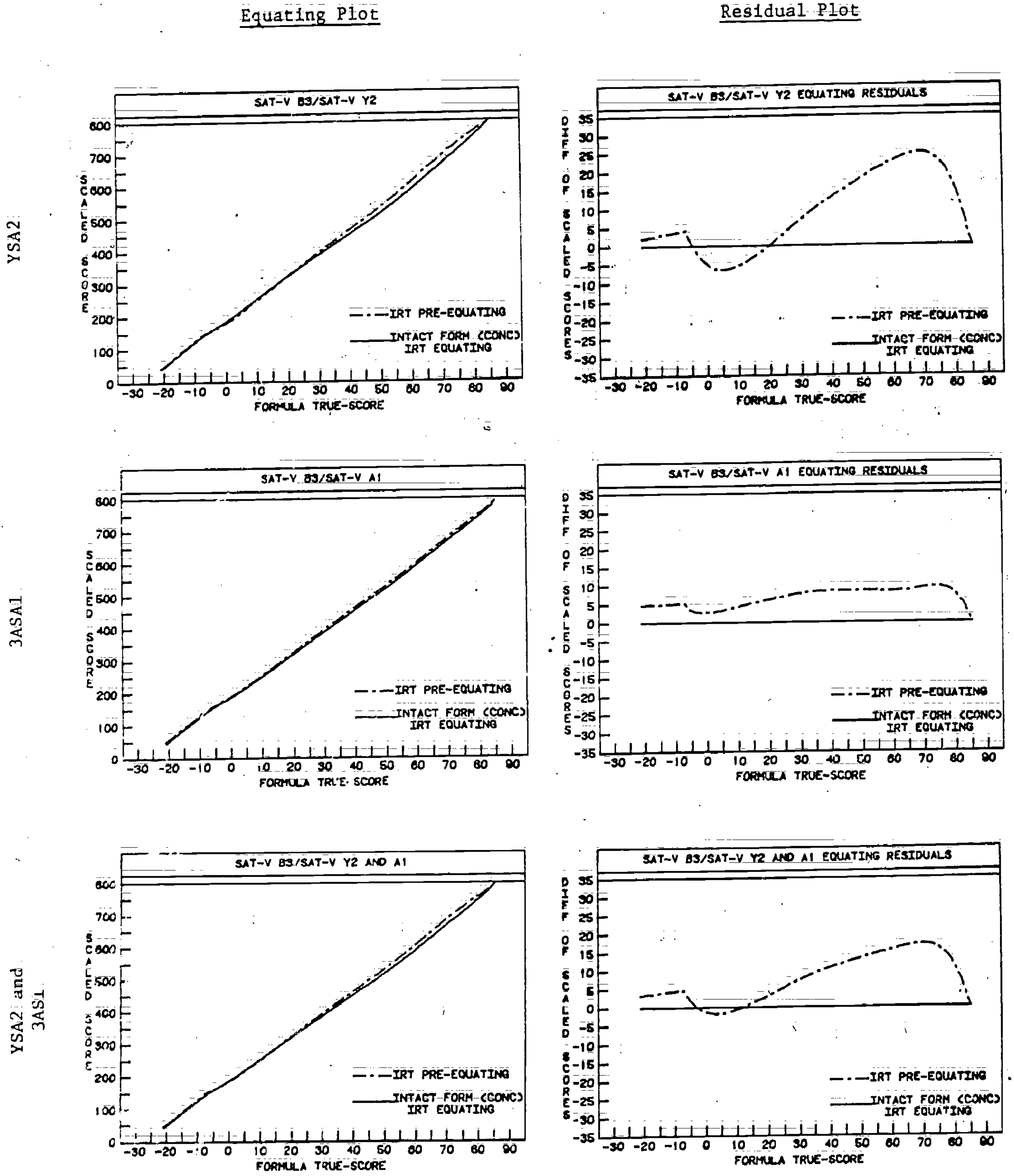


Figure 9: SAT-verbal Form 3BSA3 equated to SAT-verbal Form YSA2, Form 3ASA1, and Forms YSA2 and 3ASA1 - Plots of 1) IRT pre-equating raw to scaled score transformation compared to intact final form concurrent IRT equating raw to scaled score transformation, and 2) differences between scaled scores (IRT pre-equating - intact final form concurrent IRT equating) resulting from the equatings.

Table 5

Scaled Score Summary Statistics Resulting from Application of Five Equating Methods
SAT-verbal Sections of Form 3BSA3

| Form | N | | IRT Intact Form (Direct Link) | IRT Intact Form (Calibration System) | Intact Form Linear | IRT Intact Form (Concurrent Run) | IRT Pre-equating |
|---|---|---|---|---|---|---|---|
| 3BSA3 | 253,354 | M | 430.25 | 430.42 | 431.42 | 431.54 | 440.39 |
| | | S.D. | 105.99 | 105.57 | 106.53 | 105.86 | 110.55 |

The only other possible explanation for the differences between the Form 3BSA3 pre-equating and intact form comparison equating results has to do with the quality of the parameter estimates for the 85 3BSA3 items when they appeared in pretest form. In order for the equatings to be as discrepant as they are, the pretest and final form parameter estimates for certain of the items must be quite different. The following method was used to compare these two sets of parameter estimates in an attempt to locate those items for which the pretest parameter estimates were problematic. A mean absolute difference and a mean signed difference between the item response functions for each item, where the functions were generated using the pretest and the final form parameter estimates, were obtained. Using all individuals in the sample taking Form 3BSA3 when calibrated as an intact final form, the absolute difference and the signed difference between the item response functions for each person (i.e., value of θ) were obtained and then averaged across people. Items having the largest mean absolute difference and signed absolute difference values were then located. The above analysis was also done for the two sets of Form 3ASA3 item parameter estimates so that the discrepancies between parameter estimates for 3ASA3, where the pre-equating results were more than acceptable, could be compared to the 3BSA3 discrepancies.

Using the mean absolute and mean signed differences between the item response functions as criteria for selection of problematic items, thirteen items from 3BSA3 and twelve from 3ASA3 were identified. Upon inspection of these two subsets of problem items, they were found to differ considerably in characteristics. Of the thirteen items identified for Form 3BSA3, eleven were reading comprehension items. Of the eleven, four were based on the same passage and three on another passage. The remaining four reading comprehension items were single items based on four different passages. Of the twelve items identified for Form 3ASA3, four were reading comprehension items (two from one passage, the other two from two other passages different from the first), four were antonym items, three were analogies, and one was a sentence completion item. Of the thirteen 3BSA3 items identified, twelve were more difficult when given in a pretest than in the final form. Of the twelve 3ASA3 items, seven were more difficult when given in a pretest and five when given as part of the intact final form.

Upon closer inspection of the eleven reading comprehension items from 3BSA3 exhibiting large absolute and signed differences in item response functions, it was found that nine of these items came from pretests in which the passage they were linked to was located in the final position of the pretest section. For all but one of these items, the item as it appeared in the pretest was more difficult, sometimes considerably more, than when it appeared in the final form. For the lone exception, a word was deleted from the correct response (the only such occurrence on either 3ASA3 or 3BSA3) between the time when the item was given in a pretest and the time it appeared in the final form. This word also appeared as a key word in the text of the passage and it may be hypothesized that it acted as a "clue" to the correct response, thus explaining the increase in difficulty upon removal of the word from the correct response when the item appeared in the final form. Figure 10 contains plots of the item response functions based on pretest and

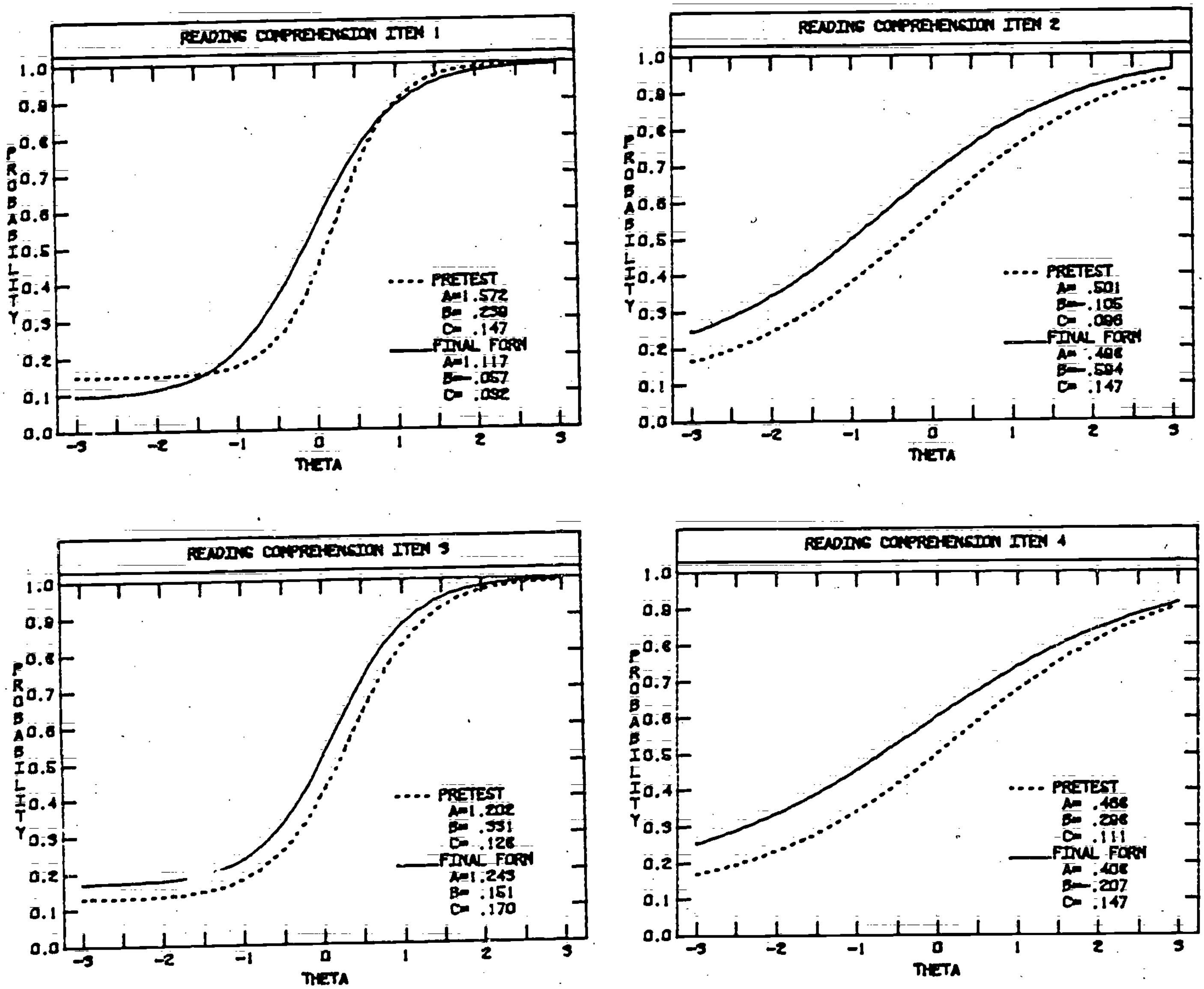


Figure 10: Plots of item response functions based on pretest and intact final form item parameter estimates for thirteen problematic Form 3BSA3 items.

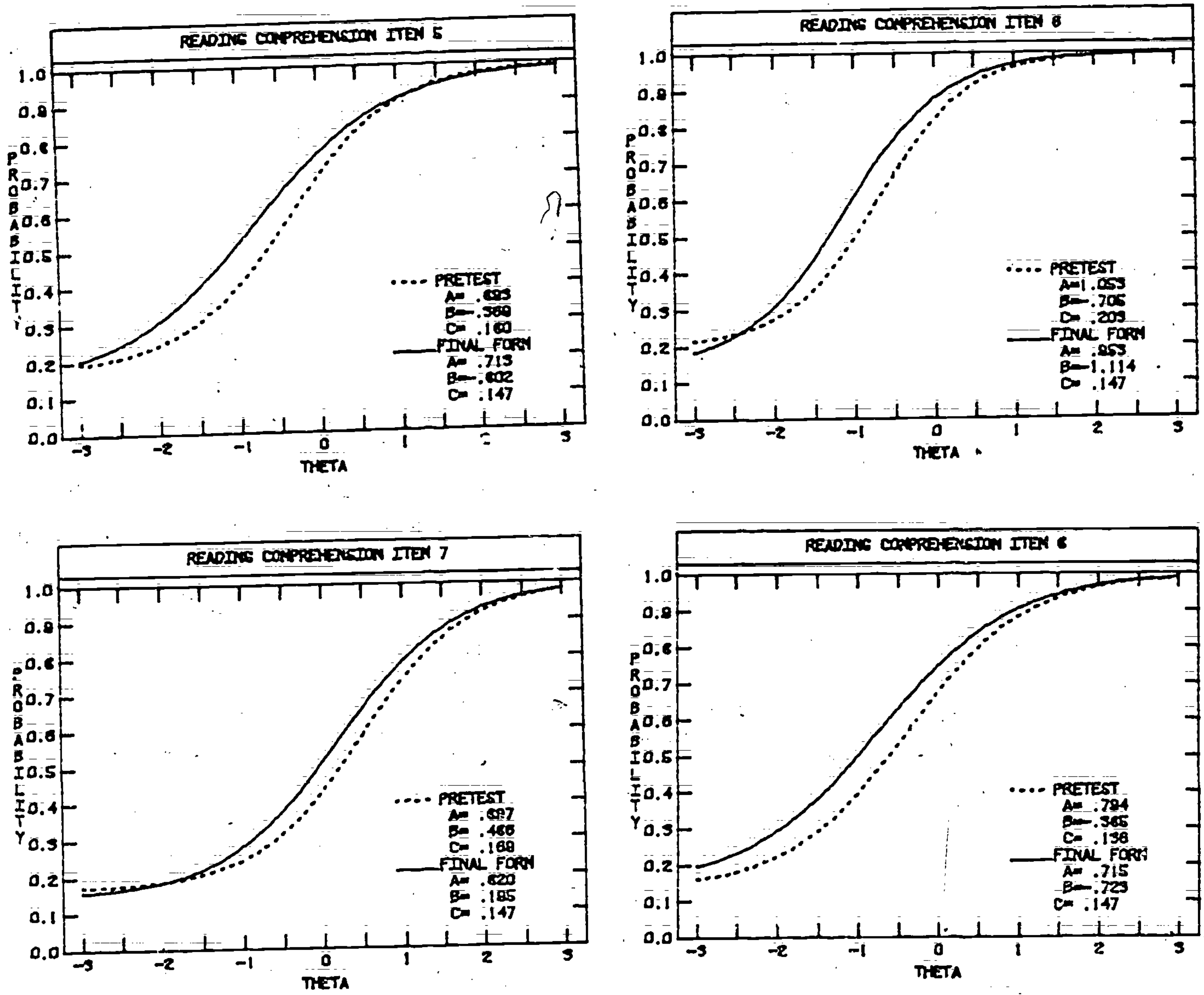


Figure 10: Plots of item response functions based on pretest and intact final form item parameter estimates for thirteen problematic Form 3BSA3 items.

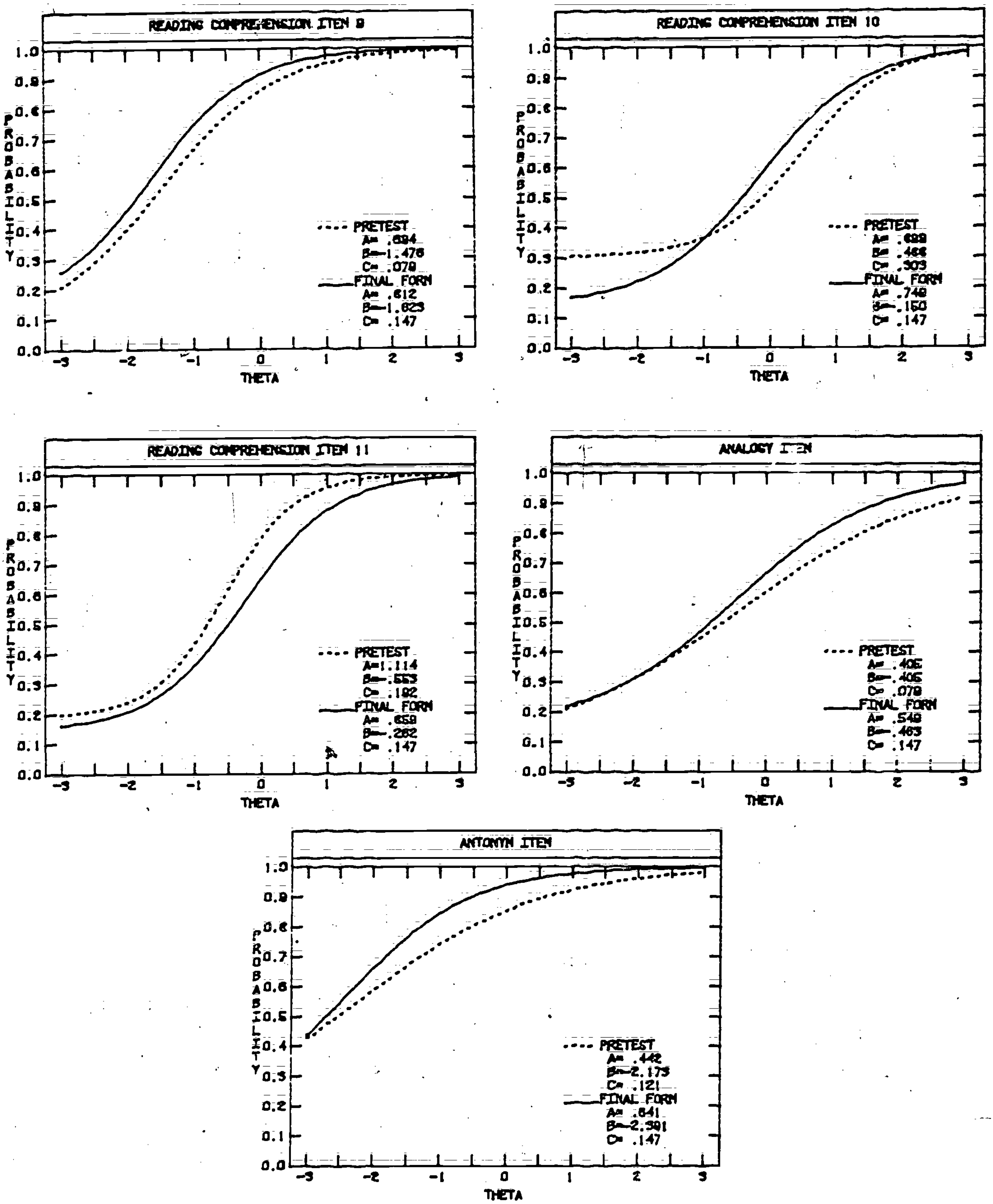


Figure 10: Plots of item response functions based on pretest and intact final form item parameter estimates for thirteen problematic Form 3BSA3 items.

intact final form parameter estimates for the thirteen problematic 3BSA3 items, identified by item type. For the reading comprehension items (numbered 1-11), items numbered 1-4 are all based on the same reading passage, as mentioned earlier, and items numbered 5-7 are based on the other passage discussed. Reading comprehension item number 11 is the item in which the word was deleted from the correct response between when the item was given in pretest and in final form. The remaining two problematic items are presented after the reading comprehension items; one of the items is an analogy item and the other is an antonym item.

Upon inspection of the four problematic reading comprehension items identified in Form 3ASA3, it was determined that, exactly like the situation for Form 3BSA3, the items were located in pretests where the passage they were linked to was located last in the pretest section. All four items were also more difficult when they appeared in the pretest section than in the final form.

On the basis of the above data, it may be hypothesized that either something approaching a "fatigue factor" is being exhibited in the responses of candidates to reading comprehension item: based on passages located at the end of pretest sections or, because of lack of time, random responses are being supplied by these candidates to certain of the questions based on this last passage. In either case, the items are more difficult in the pretest than they are in the final form, where due to passage location, a fatigue factor or the supplying of random responses is not occurring. The data on the reading comprehension items from both 3BSA3 and 3ASA3 are consistent with this statement.

If the above is happening to pretest reading comprehension items based on passages located at the end of pretest sections, one might be concerned about whether there are large discrepancies in parameter estimates between pretest and final form for reading comprehension items in the intact final form based on passages at the end of the SAT-verbal 45 item and 40 item sections. The 45 item SAT-verbal sections do not end with reading comprehension items, but the 40 item sections do. For 3ASA3, the passage upon which the last set of reading comprehension items (items 36-40) were based was also located in the final position in the pretest. Two of the five items demonstrated discrepancies large enough to be included in the overall set of twelve items discussed earlier. For 3BSA3, the passage upon which the last set of reading comprehension items (items 36-40) were based was not located at the end of the pretest section. It was, however, the only reading comprehension passage in the pretest, and one of these last five items (36-40) in 3BSA3 did exhibit large discrepancies in pretest-final form parameter estimates. Thus it would appear that while the outcome in terms of parameter estimate discrepancies for reading comprehension items located at the end of SAT-verbal sections is not as clear cut as for comparable pretest reading comprehension items, there is still cause for concern.

The effect on equating of having, in particular, reading comprehension pretest item difficulties estimated to be lower than they are when estimated on intact final form data is predictable, and demonstrated in the Form 3BSA3 pre-equating results. If the same items are more difficult in the first "test" (made up of pretest items) than in the second (made up of the items in the intact final form), then the same raw score on both "tests" should result in a higher scaled score on the first "test" than the second. This appears to be exactly what is happening with the 3BSA3 pre-equating results. For 3ASA3, on the other hand, there is more of a balancing effect of the discrepancies between pretest and intact final form parameter estimates and the result is that the pre-equating and the intact final form comparison IRT equatings more closely coincide. One might still be concerned, however, about the fact that any discrepancies at all showed up between the pretest and intact final form parameter estimates, particularly for 3ASA3, where the problem with final passage reading comprehension items was minimal. A study by Cook, Eignor, and Petersen (1982), examining the stability over time of intact final form SAT-verbal (and other testing data) parameter estimates, can be used to address this issue. The magnitudes of the discrepancies found in the Cook, et al, (1982) study, based on the same intact final form items given on two occasions, were of the magnitudes of the discrepancies found in this study. Hence, the lack of parameter invariance demonstrated by certain of the items in this study may not be such a serious cause for concern; this will be discussed further in the conclusions section.

Additional Equating Results

In concluding this section, one other noteworthy result should be mentioned; this follows from a comparison of the intact form calibration system IRT equating to the intact form direct link and intact form concurrent IRT equatings. It would appear, based upon the equatings done, that the equating is more than adequate when done through the indirect linking of the new and old forms used for equating via the overall calibration system. That is, even though in this situation the forms to be equated are linked indirectly through intervening LOGIST runs, and parameter estimates placed on a scale defined by the ability distribution of the sample taking a form not used in the equatings, the quality of the equatings are comparable to those resulting from either linking the new and old forms directly (direct link equating) or calibrating all data concurrently so that new and old form parameter estimates are automatically on the same scale. This has important implications for the future construction of a large pool of calibrated items and test forms to be used in intact final form IRT equating of SAT-verbal.

## Conclusions

The results of pre-equating the two forms of SAT-verbal reported on in this study, when compared to the intact final form IRT equatings, varied considerably, ranging from acceptable for Form 3ASA3 to only marginally acceptable for Form 3BSA3. Reasons for the inferiority of the Form 3BSA3 pre-equating results, having to do with the location of reading passages and reading comprehension items at the end of pretest sections, have been advanced and discussed. The overall results of the pre-equating of the two forms of SAT-verbal were deemed sufficiently promising that an investigation of pre-equating two forms of SAT-mathematical is presently being undertaken. Data from the second study, when considered with the data from this study, should supply information necessary for the consideration of IRT pre-equating of the

SAT on a regular operational basis. The results reported here also have clear implications for changes in test development practice, having to do with the positioning of pretest and final form reading comprehension items and making "minor" changes in the wording of items between pretest and final form, if pre-equating the SAT-verbal section is to become a reality.

On a more general level, the results of this study indicate that the IRT item parameter estimates generated for certain items when given in pretest form did not remain invariant when given in intact final form. Based on the results of recent studies, particularly Cook, Eignor, and Petersen (1982), parameter invariance for all items in a test form would not be expected to be the case. The real issue is whether the lack of parameter invariance is serious enough to cause one to dismiss the use of item response theory for the particular application of concern. The application in this study is pre-equating and the results, particularly as they pertain to SAT-verbal Form 3ASA3, suggest that even though there is some lack of parameter invariance, IRT pre-equating may be a reasonable equating method for the SAT-verbal section.

## References

Angoff, W. H. Test reliability and effective test length. Psychometrika, 1953, 18, 1-14.

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education, 1971.

Bejar, I. I., and Wingersky, M. S. A study of pre-equating based on item response theory. Applied Psychological Measurement, 1982, 6, 309-325.

Cook, L. L., and Eignor, D. R. Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia, 1983.

Cook, L. L., and Petersen, N. S. Item response theory equating for the SAT: New designs and directions. Proposal submitted to College Board — ETS Joint Staff Research and Development Committee, 1982.

Cook, L. L., Eignor, D. R., and Petersen, N. S. A study of the temporal stability of IRT item parameter estimates. A paper presented at the annual meeting of AERA, New York, 1982.

Kingston, N. M., and Dorans, N. J. The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test. RR-82-12. Princeton, NJ: Educational Testing Service, 1982.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1977, 14, 117-138.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Petersen, N. S., Cook, L. L., and Stocking, M. L. IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, in press.

Stocking, M. L., and Lord, F. M. Developing a common metric in item response theory. RR-82-25-ONR. Princeton, NJ: Educational Testing Service, 1982.

Wingersky, M. S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia, 1983.

Wingersky, M. S., Barton, M. A., and Lord, F. M. LOGIST V user's guide. Princeton, N.J.: Educational Testing Service, 1982.